

Original Articles

Estimating soil organic carbon stocks using different modelling techniques in the semi-arid rangelands of eastern Australia



Bin Wang^{a,*}, Cathy Waters^b, Susan Orgill^a, Annette Cowie^c, Anthony Clark^b, De Li Liu^a, Marja Simpson^b, Ian McGowen^b, Tim Sides^a

^a NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, NSW 2650, Australia

^b NSW Department of Primary Industries, Orange Agricultural Institute, NSW 2800, Australia

^c NSW Department of Primary Industries, Trevenna Rd, Armidale, NSW 2351, Australia

ARTICLE INFO

Keywords:

Soil organic carbon stocks

Land fractional cover data

Remote sensing

Machine learning

Genetic algorithm

Stepwise regression

ABSTRACT

Soil organic carbon (SOC) is pivotal for biological, chemical and physical processes and provides vital information on changes in soil fertility and land degradation. Rangelands, accounting for about 81% of Australian land area, are significant carbon (C) stores and small increases in soil C sequestration over such a vast area represents a considerable climate change mitigation opportunity. Efficient modelling techniques to evaluate the potential to increase rangeland SOC stocks are vitally important to assess their role in the global carbon cycle and quantum abatement. This study aimed to evaluate boosted regression trees (BRT) and random forest (RF) models in predicting SOC stocks from available continuous remotely sensed variables using two feature selection techniques. Dominant variables that affect SOC stocks in the rangelands were also identified. Using field-based measurements of SOC stock collected from 564 data points across the study area and 28 GIS-based environmental variables including climate, topography, radiometry, vegetation and land fractional cover data, we employed stepwise regression (SR, linear approach) and genetic algorithm (GA, nonlinear approach) to select the most informative variables. These selected predictors were then used to train the BRT and RF models. In all, four models were evaluated: BRT using SR (SR_BRT); RF using SR (SR_RF); BRT using GA (GA_BRT) and RF using GA (GA_RF). In addition, BRT using all predictors (All_BRT) and the RF using all predictors (All_RF) were used as benchmarks to test the performance of the four models. Of the field-based data, 75% were used to train the model ("calibration dataset") and the remaining 25% were used to validate the prediction of SOC stocks ("validation dataset"). The results indicate that the RF exhibited a better performance in predicting SOC stocks than the BRT regardless of input variables. In addition, we verified that feature selection for both machine learning techniques is necessary for estimating SOC stocks because they can increase accuracy and save time. The GA_RF was the most reliable method to predict SOC stocks, with the lowest root mean square error (RMSE) and the highest R^2 values (7.44 Mg C ha⁻¹ and 0.48, respectively), suggesting that the method of using GA-RF to generate a predictive model from measured data and remotely-sensed variables may provide a cost effective alternative to direct sampling to predict SOC stocks in the semi-arid rangelands of eastern Australia. The important variables for explaining the observed SOC stocks were rainfall, elevation, Prescott index (PI, a measure of water balance), and land fractional cover (bare ground fraction). The approach proposed here can be extended in areas where field observed data is scarce (e.g. rangelands) to produce more detailed information about SOC stocks. As such, the results of our study are of particular importance in Australian rangelands to provide a statistical and theoretical basis for producing digital SOC stock maps based on readily available remotely-sensed data, with potential for use in similar rangelands conditions internationally.

1. Introduction

Approximately half of the global land mass is occupied by rangelands including grasslands, shrub lands, deserts and tundra, which contain more than a third of above- and below-ground carbon (C)

reserves (Schuman et al., 2002). Rangelands extend across low rainfall and variable climates, accounting for about 81% of Australian land area (<http://www.environment.gov.au/land/rangelands>) (Allen et al., 2013). It is estimated that Australia's rangeland soils contain between 34 and 48 Gt of carbon, representing a sequestration potential of 78 Mt

* Corresponding author.

E-mail address: bin.a.wang@dpi.nsw.gov.au (B. Wang).

<https://doi.org/10.1016/j.ecolind.2018.01.049>

Received 29 August 2017; Received in revised form 24 January 2018; Accepted 24 January 2018

Available online 02 February 2018

1470-160X/ © 2018 Elsevier Ltd. All rights reserved.

Table 1

Summary of field surveys in south eastern Australia where SOC stocks (0–0.30 m) were compared under different land use and management. Note: there is some overlap in these studies due to the same data being used in multiple papers for different comparisons. Abbreviations include: average annual rainfall (AAR), soil pH in water (pH_w), Colwell phosphorus (Col P) and vapour pressure deficit (VPD).

Agricultural system and management comparison	Sites (total)	Significant difference (Y/N)	Comparisons where difference in OC stock were significant	Other significant factors influencing OC stock	Location and reference
<i>Rain-fed crop and pasture</i> Crop: continuous, rotation, pasture croppingPasture (perennial) : low vs high nutrient input Grazing: continuous vs rotational Crop: continuous, rotation, pasture cropping Pasture: introduced perennial	354	N		Climate (AAR), elevation, silica content (texture)	NSW Badgery et al. (2013)
Pasture: introduced perennial vs native, perennial vs annual, phosphate applied vs no phosphate, pasture cropping vs control (crop or pasture) Grazing: continuous vs rotational grazing Crop: continuous, rotationPasture: introduced perennial, native (voluntary mixed pastures)	223	Y	Pasture > crop (continuous and rotation)	Climate (AAR), soil properties (silica content, pH _w , gravel, Col P) and pasture composition	NSW Badgery et al. (2014)
	44	Y	Phosphate applied > no phosphate (pasture)	Climate (AAR), previous land use	NSW and Victoria Chan et al. (2010)
	200	Y	Pasture > crop ¹	Climate, topography and soil properties (EC, pH, Col P, CEC, exchangeable sodium (Na), potassium (K), calcium (Ca), magnesium (Mg) and aluminium (Al))	NSW Davy and Koen (2013)
	615	N		Climate (AAR VPD, temperature) and soil properties (silica content)	Victoria (Robertson et al., 2016)
Crop: continuous, rotation, stubble and tillage management Pasture: introduced perennial vs annual or mixed Grazing: continuous vs rotational Crop: continuousPasture: introduced perennial (tropical) and native	145	Y	Pasture > crop ²	Climate, soil type and duration of land use	NSW Schwenke et al. (2013)
Crop: rotation (with cultivation) Pasture: introduced perennial, native pasture Remnant woodland	20	Y	Woodland > non-woodland	Soil type	NSW Wilson et al. (2011)
Crop: continuousPasture: introduced perennial (grass) or lucerne Grassy woodland	22 ³	Y	Non-crop > crop	Soil type	NSW Young et al. (2005)
<i>Rain-fed crop and pasture, irrigated crop (cotton) or pasture</i> Pasture: introduced perennial (species comparison), native, inputs: irrigated with dairy effluent vs mineral fertiliser Native forest	20 ³	Y	Introduced > native pasture Introduced pasture irrigated with dairy effluent > mineral fertiliser	Climate (AAR), previous land use	NSW Chan and McCoy (2010)
Crop: continuous, rotation, cotton (irrigated) Pasture: introduced perennial, native Grazing: low intensity, continuous, rotational Wooded: native forest, softwood plantation	1401	Y	Native forest and grass > introduced pasture $\cong$ grazing systems > softwood plantation $\cong$ crop rotation $\cong$ minimum till > cropping (tillage) > irrigated cotton $\cong$ continuous cropping	Climate and soil properties	NSW Hobley et al. (2015)
Crop: rotation, tillage and stubble management, cotton (irrigated) Pasture: introduced perennial, native Grazing: continuous vs rotational Organic amendments	780	Y	Pasture and crop rotation > continuous crop	Climate (AAR), longitude, elevation, soil pH and inputs (N fertiliser, soil conditioner, organic amendments)	NSW Rabbi et al. (2014)

¹ Only significant in Slopes region > 500 mm AAR (Davy and Koen, 2013).

² Only significant on Chromosols (Schwenke et al., 2013).

³ Comparison 0–0.20 m (Chan and McCoy, 2010; Young et al., 2005).

C per year for small increases of about 0.25% per year (Keating et al., 2009). Soil organic carbon (SOC) is also an indicator of soil fertility. Therefore, accurate assessment of the distribution and changes in SOC stocks is essential to manage and enhance this resource.

Identifying the factors that influence SOC variability across different landscapes using statistical methods is of prime importance in SOC quantification and modelling potential change given its role in climate change abatement and in meeting emissions reduction targets globally. The estimation of SOC stocks using statistical models underpinned by environmental variables (such as climate, soil properties and topography) and land use has been demonstrated in a number of studies (Akpa et al., 2016; Bonfatti et al., 2016; Ließ et al., 2012; Razakamanarivo et al., 2011; Schillaci et al., 2017). In south eastern Australia, field surveys have demonstrated a significant influence of environmental variables on SOC stocks in agricultural systems (Table 1), and little or variable influence of land management (Rabbi et al., 2014, 2015). However, few studies have examined the role of environmental variables on SOC stocks in Australian rangelands. Environmental factors such as climatic variables including rainfall and temperature affect SOC storage by influencing the rate of organic matter (OM) input and decomposition and the quantity and quality of OM cycled through terrestrial systems (Akpa et al., 2016). Similarly, soil properties such as clay content, pH and nutrient availability influence the microbial degradation of OM, and therefore the stability of OC in soil (Orgill et al., 2017a; Six et al., 2002; Stewart et al., 2007). Topography affects SOC stock through its effects on soil depth, soil moisture and the redistribution of soil particles in the landscape. For example, Davy and Koen (2013) found that the most important environmental factors in predicting SOC stocks in south-eastern Australia were topographical (i.e. slope) and physico-chemical (clay mineralogy, CEC and pH of soil) variables at both the catchment and subregional scale.

Numerous statistical prediction models have been developed to correlate environmental variables and SOC stocks. Most commonly, linear regression methods have been used for relating SOC to ancillary variables (Thompson et al., 2006; Zhao et al., 2014) and typically these methods are simple and straightforward to apply, use and interpret (Thompson et al., 2006). While multiple linear regression assumes the relationship between soil properties and environmental variables are linear, in reality, these relationships are usually complex and non-linear. Recent studies have used techniques adapted from data mining and machine learning methods to account for the relationships between soil attributes and environmental variables (Adhikari et al., 2014; Hengl et al., 2015; Kuang et al., 2015; Sreenivas et al., 2016; Were et al., 2015). Several studies demonstrate that machine learning algorithms are more accurate than statistical methods such as step-wise linear regression, principal component regression and partial least squares regression (Guo et al., 2015; Mouazen et al., 2010), especially when the feature space is more complex. For example, Bou Kheir et al. (2010) successfully used classification and regression trees (CART) to model distribution patterns of SOC with remote sensing variables in Denmark and Were et al. (2015) compared the performance of support vector regression (SVR), artificial neural networks (ANN) and random forest (RF) for predicting and mapping SOC across an Afrotropical landscape. Of these techniques, tree models such as the boosted regression tree (BRT) and RF have been widely used to map the spatial distribution of SOC. They have advantages including fewer parameters and the ability to investigate non-linear and hierarchical relationships between the predictors and the response (Everingham et al., 2016). However, a major limitation of these machine learning approaches is that they only account for the relationships between SOC and environmental factors and ignore the influences (that is, spatial autocorrelation) of neighbouring observed data when predicting the spatial distribution of SOC (Guo et al., 2015).

Although various machine learning algorithms have been widely used, it is still difficult to produce a robust model to predict SOC stock

due to high levels of SOC variation and complex relationships with environmental variables. A number of other studies have considered factors to include in order to improve model performance in predicting soil properties (e.g. SOC stocks) (Guo et al., 2015; Ließ et al., 2016; Xie et al., 2015). These studies demonstrate the selection of predictor variables (i.e. feature selection) to elucidate the most relevant or informative input variables to minimize errors and develop the most robust models. The process of selecting the most informative subset of variables is necessary, to alleviate the impacts of the “curse of dimensionality”, enhance generalisation capacity, speed up the machine learning process and increase model interpretability (Rodriguez-Galiano et al., 2012; Taghizadeh-Mehrjardi et al., 2016; Xie et al., 2015). Among the numerous variable selection methods available for SOC stocks, genetic algorithms (GA) have been demonstrated to have superior performance in soil type recognition as well as other input selection studies (Huang and Wang, 2006; Welikala et al., 2015). Taghizadeh-Mehrjardi et al. (2016) used GA to select the most informative subset of variables in determining SOC stocks with various machine learning methods in the Baneh region, Iran. However, there has been no published work to date reporting how much improvement can be achieved using GA feature selection in explaining SOC stock in the rangelands.

Recent technological and methodological advancements in the field of remote sensing provide new opportunities for the use of spectral indices derived from high quality satellite imagery stacks (e.g., MODIS and Landsat) for digital soil mapping (Maynard and Levi, 2017; Minasny et al., 2013; Were et al., 2015). Spectral indices can provide useful information on soil properties indirectly through the use of vegetation indicators (Maynard and Levi, 2017). A number of studies focused on modelling SOC have incorporated vegetation indices to predict SOC stocks (Bou Kheir et al., 2010; Vågen et al., 2016) as vegetative cover can reflect ecosystem productivity and OM input to soil (Román-Sánchez et al., 2016). For example, the Normalized Difference Vegetation Index (NDVI), the most commonly used band ratio in ecological environment research, has been widely applied in studies of SOC. Page et al. (2013) found that NDVI was one of the most important variables affecting SOC stocks across grain-cropping regions of Queensland (Eastern Australia). Bou Kheir et al. (2010) used the Normalized Difference Wetness Index (NDWI) in their classification-tree based models to predict SOC in Denmark, however it was not a major contributor to the variation in SOC. Yang et al. (2008) predicted the spatial distribution of SOC stocks using a satellite-based enhanced vegetation index (EVI) across the Tibetan grasslands. They reported that EVI explained 50–66% of the variation in SOC stock in their prediction models. Recent studies in the semi-arid rangelands have shown clear relationships between ground cover (perennial grass and litter cover) and SOC stock (Orgill et al., 2017b; Waters et al., 2015, 2016). These relationships suggest suitable satellite-derived covariates such as fractional cover data may be useful in the estimation of SOC stocks in these semi-arid environments.

In this context, the objective of this study was to evaluate alternative methods to derive predictions of SOC stocks from directly-measured and remotely-sensed variables. Specifically, this study aimed to (1) evaluate feature selection techniques to identify the dominant variables that affect SOC stocks and (2) compare RF and BRT methods to determine the most reliable and accurate model to predict SOC stocks in the surface 0.30 m of the soil profile in the semi-arid rangelands of eastern Australia.

2. Materials and methods

2.1. Study area

The model training area (referred to as the study area), is located in western New South Wales, Australia (Fig. 1). It lies between a latitude of 29.64° and 32.28° South and a longitude of 145.54° and 146.06° East.

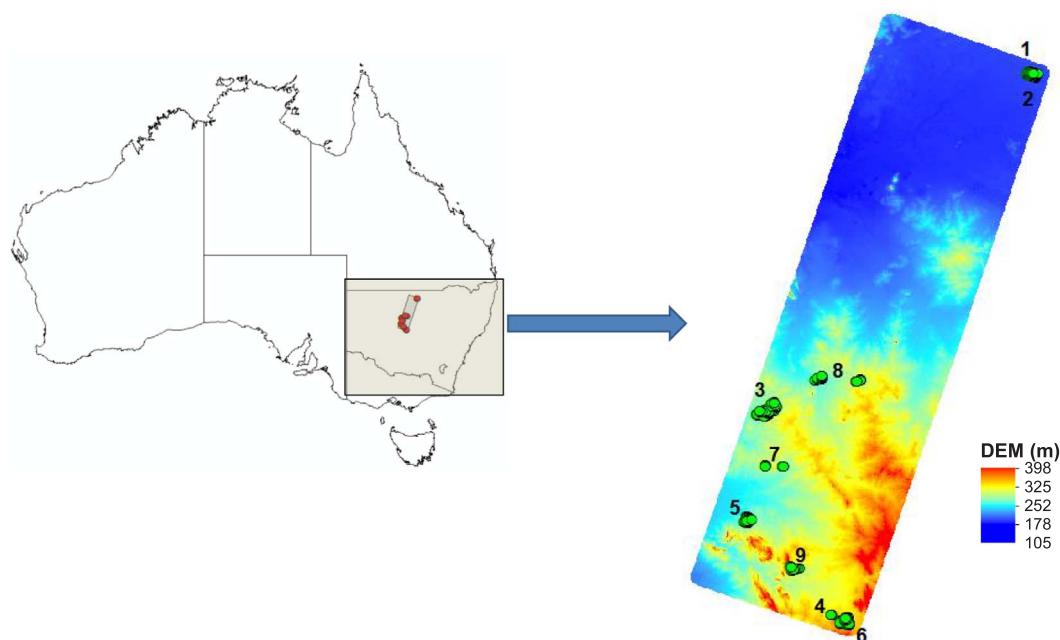


Fig. 1. The study area in the semi-arid rangelands of eastern Australia. A total of 564 data points (green dots) were collected from nine locations (outlined in Table 2).

The dominant land use is extensive grazing and the major soil types (IUSS-Working-Group and WRB, 2015) are Vertisols (in the northern part of the study area) and Lixisols (in the south). The climate is classified as BSh (arid-steppe-hot arid) in the Köppen-Geiger climate classification (<http://koeppen-geiger.vu-wien.ac.at/present.htm>) with an average annual rainfall and temperature, 379 mm and 18.9 °C respectively. The terrain is generally flat with the elevation ranging from 105 to 398 m. The vegetation in the study area is dominated by small patches of shrubs, scattered trees and large areas of open grasslands.

2.2. Soil sampling and analysis

A total of 564 data points within the study area were used to calibrate and validate the models. Each data point represents a soil core (to at least 0.30 m), the coordinates of which were randomly selected prior to field sampling (Table 2). Soil cores were carefully collected using a hydraulically driven core sampler with a 40 mm inner core diameter to a depth of at least 0.30 m and divided into four soil depths; 0–0.05 m, 0.05–0.10 m, 0.10–0.20 m and 0.20–0.30 m. Soil OC was determined on approximately 2 g of finely ground < 2 mm soil analysed on a LECO

combustion furnace (LECO 1995) (Rayment and Lyons, 2011; Method 6B2b). Bulk density (BD) was determined on subsamples dried at 105 °C as described by Dane and Topp (2002). Results were reported as SOC g (100 g)^{−1} and BD g cm^{−3} on an oven-dry basis. More information on sampling, field measurements and soil analysis are provided in Waters et al. (2016) and Orgill et al. (2017b).

2.3. Calculation of SOC stocks

For each soil core, SOC stock for each soil layer was calculated according to Eq. (1) using SOC concentration and BD. SOC stocks from the four layers were summed to obtain the SOC stock to 0.30 m soil depth.

$$SOC_{stock} (MgCha^{-1}) = C \times BD \times D \times \left(1 - \frac{gravel [\%]}{100}\right) \quad (1)$$

where C is the concentration of soil carbon (g C (100 g)^{−1} sieved soil); BD is bulk density of the whole soil (g cm^{−3}); D is the thickness of the corresponding soil depth layer (cm); gravel [%] is the percentage of gravel in the soil sample.

Table 2

The detailed information for 564 data points collected at nine locations in the rangelands of eastern Australia.

ID	Location	Label	Number of soil cores (0–0.30 m)	Number of 0.25 m ² ground cover quadrats	Soil type WRB ¹	Refs.
1	Bokhara	BOK	40	90	Vertisol	Waters et al. (2016); Orgill et al. (2017b)
2	Tara	TARA	45	90	Vertisol	Waters et al. (2016); Orgill et al. (2017b)
3	Gilgunnia	GIL_Box/GIL_Ridge/ GIL_Yarran	120	240	Lixisol	Waters et al. (2016); Orgill et al. (2017b)
4	Etiwanda	ET	71	142	Lixisol	Waters et al. (2016); Orgill et al. (2017b)
5	Bulgoo	BUL	36	72	Lixisol	Unpublished
6	Etiwanda_Pilot	ET Pilot	102	102	Lixisol	Waters et al. (2015)
7	Box	CCC-Box (L/M/H)	45	90	Lixisol	Unpublished
8	Mulga	CCC-Mulga (L/M/H)	60	120	Lixisol	Unpublished
9	Pine	CCC-Pine (L/M/H)	45	120	Lixisol	Unpublished

¹ World Reference Base for Soil Resources (WRB) (IUSS-Working-Group, WRB, 2015).

Table 3
Descriptive statistics of soil organic carbon (SOC) stock (Mg C ha^{-1}).

Depth (m)	Min	Max	Mean	Median	SD	CV (%)
0–0.05	0.33	40.26	7.19	6.70	4.06	56.39
0.05–0.10	0.58	18.37	5.22	4.97	2.40	45.93
0.10–0.20	1.49	39.42	6.92	6.41	3.28	47.35
0.20–0.30	0.36	26.77	5.47	4.91	2.71	49.41
0–0.30	5.08	88.23	24.80	24.34	9.64	38.86

Table 3 shows the descriptive statistics of SOC stock. SOC stock in the surface 0.30 m of the soil profile was used to identify the relationship between SOC and environmental variables. The SOC stocks at 0–0.30 m ranged from 5.1 to $88.2 \text{ Mg C ha}^{-1}$, with a mean and median of 24.8 and $24.3 \text{ Mg C ha}^{-1}$, respectively. The standard deviation and coefficient of variation were 9.6 Mg C ha^{-1} and 38.9%, respectively, representing moderate variability of SOC stocks across the entire dataset.

2.4. Climate, topography and GIS-based environmental variables

The latitude-longitude coordinates were used to obtain climatic, topographic and GIS-based environmental variables for each soil data point. Climatic data, interpolated from point measurements, were obtained from SILO patched point dataset (PPD, <http://www.longpaddock.qld.gov.au/silo/ppd/index.php>) (Jeffrey et al., 2001), and 56-year (1961–2016) mean annual temperature and rainfall were calculated.

Topographic attributes are the most widely used predictors of SOC stocks, as landscape features such as elevation are important in soil formation (Behrens et al., 2010; Yang et al., 2016) and digital elevation models (DEM) are widely available. In this study, 30 m-resolution data were sourced from the Soil and Landscape grid of Australia (<http://www.clw.csiro.au/aclep/soilandlandscapegrid/>). The data were derived from the smoothed version of the national 1 arc-second resolution DEM. Ten derivatives were used, namely slope, aspect, topographic wetness index (TWI), partial contributing area (CA), MrVBF, MrRTF, plan curvature (PLC), profile curvature (PRC), Prescott index (PI) and topographic position index (TPI). All these topographic variables were downloaded from CSIRO data access portal (<https://data.csiro.au/dap/search?q=srtm&p=1&rpp=25&sb=RELEVANCE>) (Table 4).

Landsat data were obtained for each soil core location. Landsat images matched the date of soil sampling as close as possible. Landsat 8 Operational Land Imager (OLI) Nadir Bidirectional Reflectance Distribution Function Adjusted Reflectance (NBAR) data were downloaded from Australian Geoscience Data Cube (AGDC, <http://www.datacube.org.au/>). The large number of available bands from Landsat 8 makes it possible for the calculation of indices to characterise different spectral responses. Numerous vegetation spectral indices have been introduced for mapping soil properties (Poggio et al., 2013), but the selection of the best index from a large number of satellite-derived covariates for analysing soil information is one of the most important issues for the users. Of these spectral indices, the NDVI and EVI are the two most commonly used vegetation indices (independent variables) for modelling SOC stocks (Kunkel et al., 2011; Page et al., 2013; Takata et al., 2007; Wilson et al., 2017) as they indicate plant growth, specifically live green vegetation and therefore can infer net primary productivity (NPP). However, in semi-arid rangelands the effects of exposed soil, rocks and litter cannot easily be accounted for by vegetation indices. Therefore in this present study, in addition to NDVI and EVI, five spectral indices derived from Landsat Surface Reflectance were used, viz. Soil Adjusted Vegetation Index (SAVI), Modified Soil Adjusted Vegetation Index (MSAVI), Normalized Difference Moisture Index (NDMI), Normalized Burn Ratio (NBR) and Normalized Burn Ratio 2 (NBR2) (https://landsat.usgs.gov/sites/default/files/documents/si-product_guide.pdf). In addition, land fractional cover data (four bands) were downloaded from TERN AusCover (<http://data.auscover.org.au/xwiki/bin/view/Product+pages/Landsat+Seasonal+Fractional+Cover>) at a 30 m resolution. The four bands comprised Band 1-bare (bare ground and rock), Band 2-green (vegetation), Band 3-non green vegetation (litter, dead leaf and branches), and Band 4-model fitting error. These four bands were calculated using models linked to an intensive field sampling program (1500 sites) covering a broad range of vegetation types, soil and climates across Australia (Muir et al., 2011).

Filtered radiometric data were also obtained from the AGDC. The radiometrics (gamma-ray spectrometric) method is a geophysical process applied to estimate concentrations of the radioelements potassium, uranium and thorium. Detailed information about Geoscience Australia and the State and Northern Territory Geological Surveys conducted in most of Australia over the past 40 years can be found through the Geophysical Archive Data Delivery System (GADDS) (<http://www.ga.gov.au/scientific-topics/disciplines/geophysics/radiometrics>). A range of predictor variables that could be related to SOC stocks are provided in Table 4.

2.5. Modelling methodology

The overall framework, shown in Fig. 2, illustrates the procedure used in this study for feature selection and model evaluation.

2.5.1. Feature selection

The inclusion of 28 variables in predicting SOC stocks resulted in a large data set which was susceptible to extended computation times and over fitting the model due to confounding (correlated) variables and the “curse of dimensionality” (Rodriguez-Galiano et al., 2015; Rodriguez-Galiano et al., 2012). The collinearity analysis among the candidate predictors showed that vegetation indices and fractional cover data had significant relationships (Fig. 3).

We tested a linear approach (stepwise linear regression; SR) to find redundant predictors and select the smallest set of predictors giving the best linear regression results (Lima et al., 2016). We also tested a nonlinear approach (genetic algorithm; GA) to select the most informative predictor variables (Huang and Wang, 2006; Taghizadeh-Mehrjardi et al., 2016; Welikala et al., 2015). GA is a general adaptive optimization search methodology that mimics Darwinian forces of natural selection to find optimal values of a function (Huang and Wang, 2006; Taghizadeh-Mehrjardi et al., 2016). GA approach uses a set of candidate solutions (called individuals). For feature selection, the individuals are subsets of the predictors, that are represented with a binary code depending on whether or not a variable is included in the subset (Taghizadeh-Mehrjardi et al., 2016). The individuals being selected to form new solutions (offspring) are then chosen according to a pre-defined fitness function. The fitness values are calculated through a measure of model performance (root mean square error; RMSE) to assess the quality of a solution in the evaluation step. These offspring are formed using two main genetic operators (crossover and mutation) that randomly influence the fitness value. Crossover combines different features from a pair of subsets into a new subset (Welikala et al., 2015). Mutation operates by modifying some of the values through adding or removing features in a subset randomly. Offspring replaces the old population with the aim that the new population will be better than the old one. This evolutionary process is operated for many generations until termination conditions are met (i.e. iterations of the search procedure). We defined three standard parameter settings for the GA, population size of 50, crossover rate of 0.8, and mutation rate of 0.1 according to Welikala et al. (2015). SOC stock was the target variable and all ancillary variables were inputs. The process of feature selection was undertaken prior to developing SOC prediction models through the ‘Caret’ package (Kuhn, 2008) coded in R software (<https://cran.r-project.org/web/packages/caret/caret.pdf>).

Table 4
Environmental variables used in the prediction of SOC stock in the study area.

Variables	Definition and formula	Resolution
<i>Topography</i>		
Slope	The inclination of the land surface from the horizontal	30 m
Elevation	The height of a location above the Earth's sea level	30 m
Aspect	The direction in which a land surface slope faces	30 m
Topographic Wetness Index (TWI)	The relative wetness within moist catchments, but is more commonly used as a measure of position on the slope with larger values indicating a lower slope position	30 m
Partial Contributing Area (CA)	Contributing area in m ² computed using multiple flow directions on hillslopes and ANUDEM-derived flow directions in channels	30 m
MrVBF	Multi-resolution Valley Bottom Flatness	30 m
MrRTF	Multi-resolution Ridge Top Flatness	30 m
Plan Curvature (PLC)	The rate of change of aspect (across the slope) and represents topographic convergence or divergence	30 m
Profile Curvature (PRC)	The rate of change of potential gradient down a flow line and represents the changes in flow velocity down a slope.	30 m
Prescott Index (PI) ^a	Measure of water balance $PI = 0.445P/E^{0.75}$	30 m
Topographic Position Index (TPI)	Topographic position classification identifying upper, middle and lower parts of the landscape	30 m
<i>Vegetation/anthropogenic factors</i>		
NDVI	Normalized Difference Vegetation Index $NDVI = (NIR^b - R^c) / (NIR + R)$	30 m
EVI	Enhanced Vegetation Index $EVI = 2.5 * ((NIR - R) / (NIR + 6 * R - 7.5 * B^d + 1))$	30 m
SAVI	Soil Adjusted Vegetation Index $SAVI = ((NIR - R) / (NIR + R + 0.5)) * (1 + 0.5)$	30 m
MSAVI	Modified Soil Adjusted Vegetation Index $MSAVI = (2 * NIR + 1 - \sqrt{(2 * NIR + 1)^2 - 8 * (NIR - R)}) / 2$	30 m
NDMI	Normalized Difference Moisture Index $NDMI = (NIR - SWIR1^e) / (NIR + SWIR1)$	30 m
NBR	Normalized Burn Ratio $NBR = (NIR - SWIR2^f) / (NIR + SWIR2)$	30 m
NBR2	Normalized Burn Ratio 2 $NBR2 = (SWIR1 - SWIR2) / (SWIR1 + SWIR2)$	30 m
<i>Fractional cover data</i>		
Band 1	Bare ground fraction (bare ground, rock, disturbed)	30 m
Band 2	Green vegetation fraction	30 m
Band 3	Non-green vegetation fraction (litter, dead leaf and branches)	30 m
Band 4	Model fitting error	30 m
<i>Climate</i>		
Rainfall	Mean rainfall	–
Temperature	Mean temperature	–
<i>Radiometrics</i>		
Potassium	Concentrations of the radioelements potassium	100 m
Uranium	Concentrations of the radioelements uranium	100 m
Thorium	Concentrations of the radioelements thorium	100 m
Dose	Terrestrial gamma-ray dose rate	100 m

^a Referred to Gallant and Austin (2012).

^b NIR: near infrared.

^c R: red.

^d B: blue.

^e SWIR1: shortwave IR-1.

^f SWIR2: shortwave IR-2.

2.5.2. Random forest

Random forest (RF), a tree-based ensemble method was introduced by Breiman (2001) to improve regression accuracy. Initially, RF modelling draws bootstrap samples (63%) with replacement from the entire sample population in the training set to grow each tree. The bootstrap sampling method makes RF less sensitive to over-fitting compared to decision trees (Heung et al., 2014). Therefore, RF contains not a single standard regression tree but many regression trees, like a forest. It uses either categorical (i.e., classification) or continuous (i.e., regression) response variables, and either categorical or continuous predictor variables. Unlike most common methods based on machine learning, RF only needs two parameters to be set for generating a prediction model: (i) the number of regression trees to grow in the forest (n_{tree}), (ii) the number of randomly selected evidential features at each node (m_{try}). By default, the random subset size is the square root of the entire number of the predictor variables in the model.

Random forest analysis is a non-parametric algorithm that can handle non-linear and additive relationships and is used to rank the

relative importance of each predictor variable in controlling the response variable. Importance of a variable is based on the regression prediction error of the out-of-bag (OOB), which was left out of the bootstrap samples (37% of training data). It is computed as a function of change prediction error by permuting with each input variable and expressed using mean decrease in accuracy (Heung et al., 2014). In error estimation, the OOB samples were predicted by the respective trees and by aggregating the predictions and the mean square error (MSE_{OOB}) was calculated using Eq. (2)

$$MSE_{OOB} = \frac{1}{n} \sum_{i=1}^n (O_i - P_{iOOB})^2 \quad (2)$$

where n is the number of observations, P_{iOOB} is the average of all OOB predictions across all trees.

In this study, in order to optimise two parameters of n_{tree} and m_{try} , a number of experiments were conducted using different combinations of n_{tree} and m_{try} . The range of number of n_{tree} was set between 100 and 900

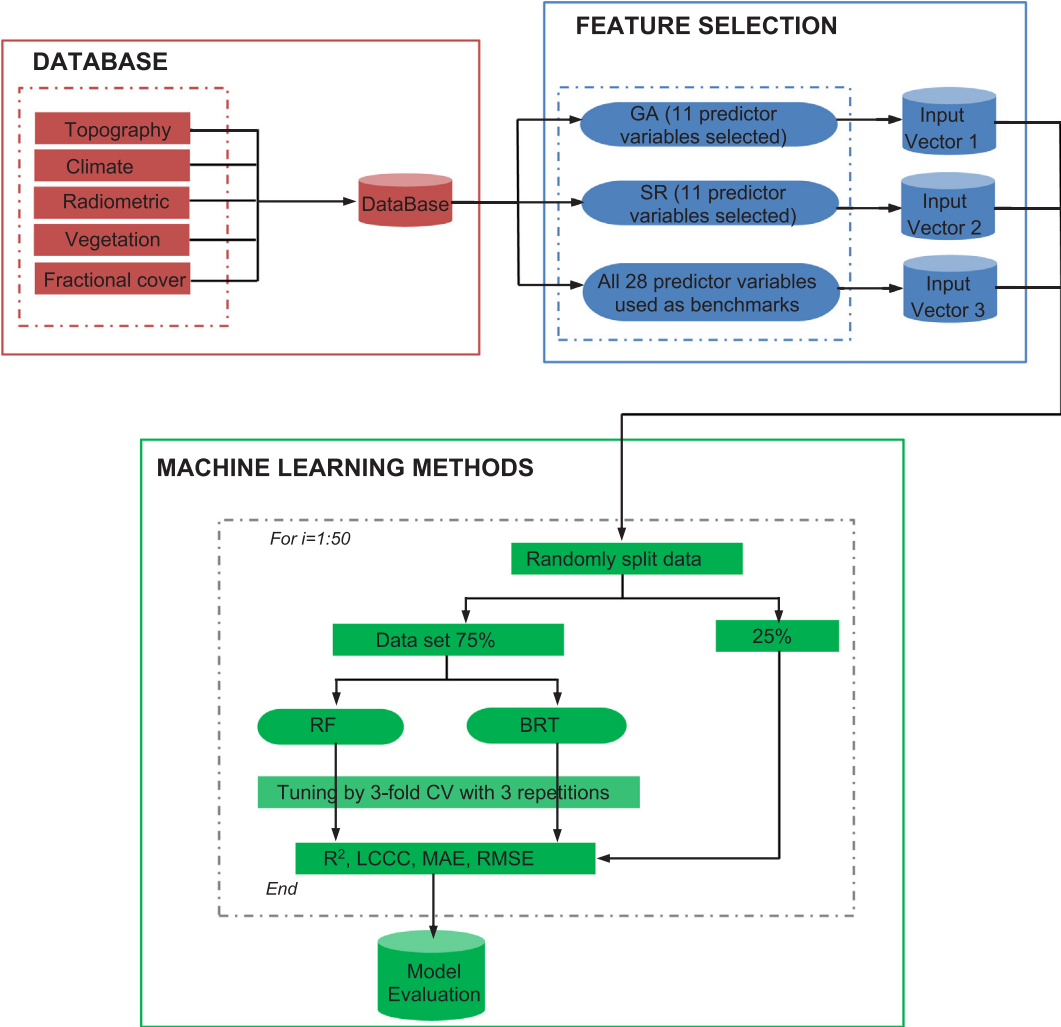


Fig. 2. Framework of procedure used, where: GA: genetic algorithm, SR: step-wise regression, RF: random forest, BRT: boosted regression tree, R^2 : the coefficient of determination, LCCC: Lin's concordance correlation coefficient, RMSE: root mean squared error and MAE: mean absolute error.

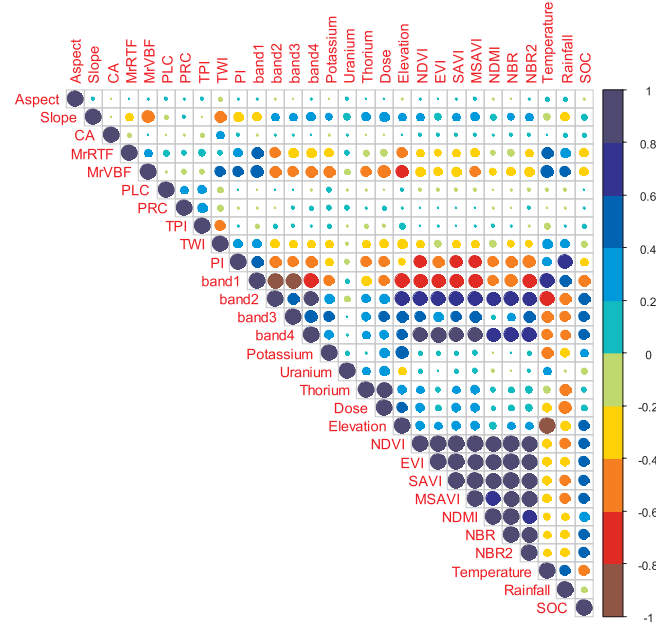


Fig. 3. Pearson correlation coefficients for the relation between SOC stock (0–30 cm) and 28 predictor variables used in this study.

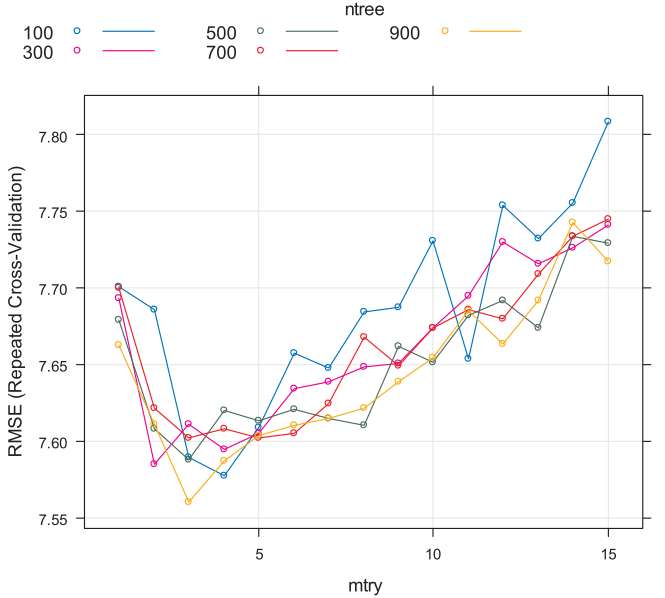


Fig. 4. The cross-validation error rates (RMSE) in tuning parameters (m_{try} and n_{tree}) for RF using training data sets.

at intervals of 200, and the number of selected evidential features between 1 and 15 at an interval of one (Rodríguez-Galiano et al., 2015). To save computing time, the training data were split into 3-folds for cross validations and the error rates for each of the 3 cross-validation partitions were aggregated into a mean error rate (Heung et al., 2014). Three replicates of the 3-fold cross-validation were conducted. The final model (optimal model) was selected when the prediction error was lowest. Fig. 4 shows the example of tuning parameters for RF using the grid research method in the ‘Caret’ package to determine the optimal parameters that produce the minimum prediction error. The cross-validation RMSE reached the least value of 7.56 Mg C ha⁻¹ when m_{try} was 3 and n_{tree} was 900 (Fig. 4).

2.5.3. Boosted regression tree

The BRT method aims to improve the predictive performance of many single models by combining many regression trees, and a boosting technique. The boosting algorithm is also referred to as a forward and stage-wise procedure in which a subset of the data is randomly selected to iteratively fit new tree models to minimize the loss function (Elith et al., 2008). A stochastic gradient boosting procedure was introduced in this process, which can improve model predictive power and minimize the risk of over-fitting. The relative importance of variables was assessed, given the number of times a variable was selected for splitting and the square improvement to each split (Ottoy et al., 2017). For this study, we fitted a BRT model via the ‘gbm’-package of R-software (<https://cran.r-project.org/web/packages/gbm/gbm.pdf>). For the BRT model, there are three main parameters: learning rate (LR) determining the contribution of each tree to the growing model, tree complexity (TC), controlling the order of interactions that can be fitted, and number of trees (NT). Similar to the RF model, in order to optimise these three parameters, a number of experiments were conducted using different combinations of NT, TC and LR. The values of TC used were 1, 3, 6, 9, 12, 15; NT varied between 100 and 2000 at 200 intervals; and the values of LR ranged between 0.001 and 0.008 at 0.001 intervals. These combinations generated an optimal BRT model through three repetitions of 3-fold cross-validation by performing a grid search using the *train* function of the ‘caret’-package. The cross-validation RMSE reached a minimum value of 7.70 Mg C ha⁻¹ when NT was 900, TC was 12 and LR was 0.003 (Fig. 5).

2.5.4. Model evaluation

We randomly selected 75% of 564 field collected soil data points for training (“calibration dataset”) and the remaining (25%) data points were used as the “validation dataset” to validate the prediction of SOC stocks (Fig. 2). To ensure model stability and increase reliability, the procedure was repeated 50 times applying a sampling with replacement method, to obtain 50 random sub-samples of the data, each one with its own calibration and validation dataset (Oliveira et al., 2012). Each calibration dataset was used to tune parameters for both machine learning methods according to the approach outlined above. Then the performance of each model with optimal parameters (identified during the feature selection approaches outlined above) was evaluated using the difference between observed and predicted response variables. To do this, four statistical indices (Yang et al., 2016) were considered: Regression Coefficients of determination (R^2) which measures the percentage of variation explained by each model; Mean Absolute Error (MAE), indicating how close the prediction is to observation; Root Mean Square Error (RMSE), measuring the overall accuracy of the prediction and Lin’s Concordance Correlation Coefficient (LCCC) measuring the level of agreement between predicted and observed values following the 45° line using the following equations:

$$R^2 = \left(\frac{\sum_{i=1}^n (O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^n (O_i - \bar{O})^2} \sqrt{\sum_{i=1}^n (P_i - \bar{P})^2}} \right)^2 \quad (3)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |P_i - O_i| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (P_i - O_i)^2} \quad (5)$$

$$LCCC = \frac{2r\sigma_o\sigma_p}{\sigma_o^2 + \sigma_p^2 + (\bar{P} - \bar{O})^2} \quad (6)$$

where P_i and O_i are the predicted and observed SOC stocks; n is the number of samples; $\bar{P}$ and $\bar{O}$ are the means for the predicted and observed SOC stocks; σ_o^2 and σ_p^2 are the variances of predicted and observed values and r is the Pearson correlation coefficient between the predicted and observed values. A good model will have LCCC and R^2 close to 1 and RMSE and MAE of almost 0.

3. Results

3.1. Feature selection

Each feature selection technique was applied to all 564 data points in order to select the most relevant predictors from the full set of predictors listed in Table 4. The linear approach revealed that 11 of the 28 environmental variables were significant for estimating SOC stocks ($P \leq 0.05$), and were therefore retained in the SR model (Table 5). These significant variables were: MrRTF, MrVBF, PI, Band 1, Band 4, potassium, uranium, elevation, NBR, temperature and rainfall. In the GA model, we used a 10-fold cross-validation with 100 iterations. The best iteration was obtained by minimizing external RMSE. The GA model with SOC stocks as the target variable and all ancillary variables as inputs, resulted in 11 predictors: MrRTF, MrVBF, PI, Band 1, Band 3, Band 4, MSAVI, TWI, TPI, elevation and rainfall. These were selected when the external RMSE reached a minimum (7.15 Mg C ha⁻¹, $R^2 = 0.45$) at iteration 87 (Fig. 6).

3.2. Assessment of model predictions

We used the selected predictors to train the RF and BRT models. Four models were evaluated: BRT using predictors selected by SR (SR_BRT), RF using predictors selected by SR (SR_RF), BRT using predictors selected by GA (GA_BRT) and RF using predictors selected by GA (GA_RF). In addition, BRT using all predictors (All_BRT) and the RF using all predictors (All_RF) were used as benchmarks to test the performance of the four models. The independent validation datasets were used to validate the model performance. The relative performance of indicators (R^2 , LCCC, RMSE and MAE) of the 50 trials using the RF and BRT based on different input features is shown in Fig. 7.

When using the validation samples, the results show that all models predicted SOC stocks moderately well based on the range of average values of R^2 from 0.42 to 0.48; LCCC ranged from 0.56 to 0.62; RMSE between 7.44 and 7.80 Mg C ha⁻¹; MAE ranged from 4.98 to 5.11 Mg C ha⁻¹. Overall, the RF model performed better than the BRT model based on the four validation measurements regardless of input features. Specifically, when using all of 28 environmental variables as input predictors, the model of All_RF had a higher R^2 (0.45) and LCCC (0.60) with a lower RMSE (7.66 Mg C ha⁻¹) and MAE (5.11 Mg C ha⁻¹) than the All_BRT model ($R^2 = 0.43$, LCCC = 0.56, RMSE = 7.79 Mg C ha⁻¹ and MAE = 5.11 Mg C ha⁻¹). In terms of the importance of the variables contributing to SOC stocks, All_RF and All_BRT showed broadly similar patterns (Fig. 8a). For example, rainfall

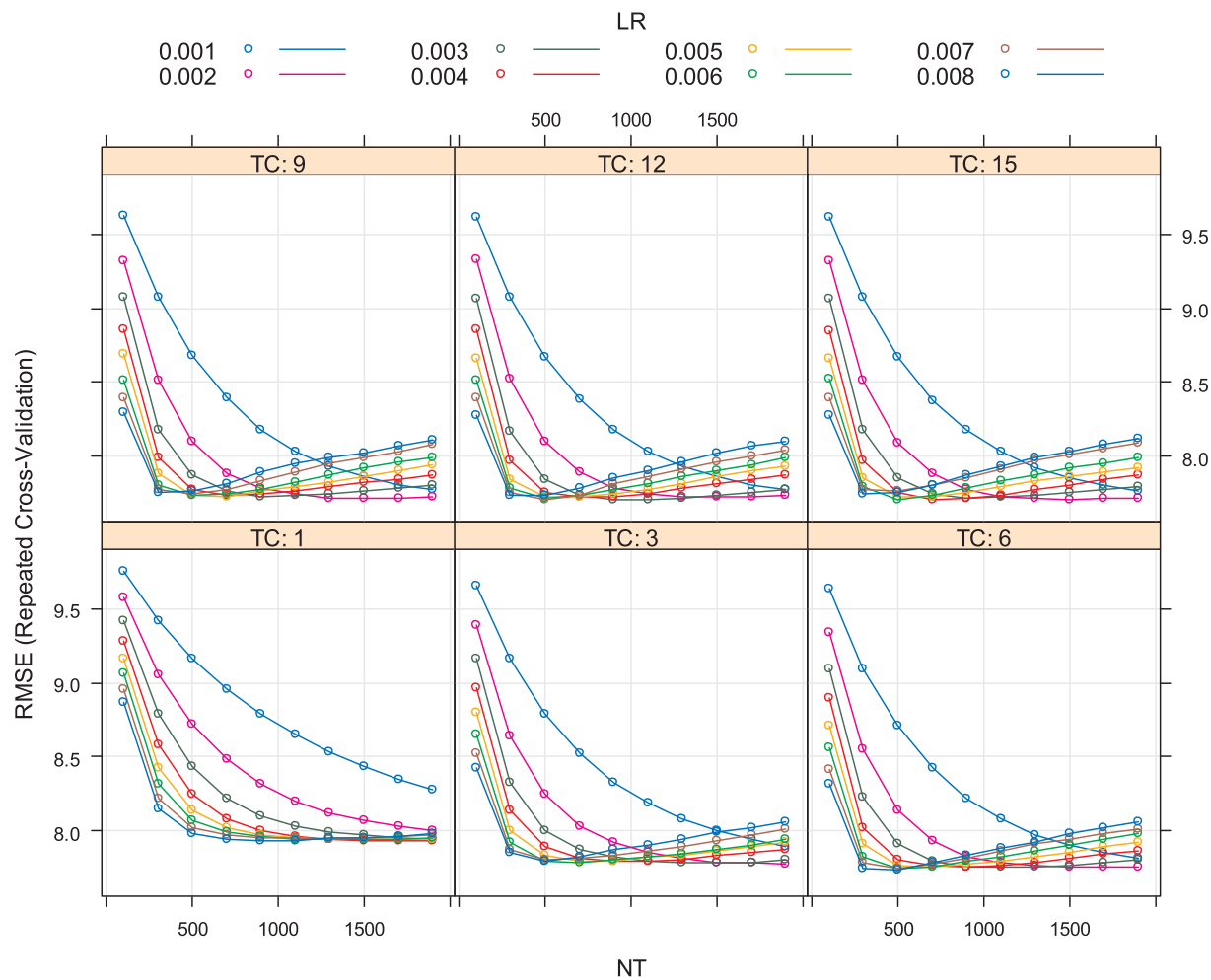


Fig. 5. The cross-validation error rates (RMSE) in tuning parameters (NT represents number of iterations; TC represents tree complexity; LR stands for learning rate) for BRT model using training data sets.

Table 5
Results of stepwise multiple regression analysis for predicting SOC stocks when 11 out of 28 alternative environmental variables were retained in the model ($P < 0.05$, $N = 564$ and $R^2 = 0.41$).

Variable	Coefficient	P value	Lower 95%	Upper 95%
MrRTF	−0.37	0.02	−0.69	−0.05
MrVBF	0.45	0.04	0.03	0.88
PI	−226.53	0.00	−339.89	−113.17
Band 1	−0.21	0.00	−0.30	−0.12
Band 4	−0.74	0.00	−1.23	−0.24
Potassium	5.90	0.00	2.44	9.36
Uranium	−4.59	0.00	−7.20	−1.99
Elevation	0.13	0.00	0.06	0.20
NBR	20.42	0.00	13.18	27.67
Temperature	5.71	0.00	2.09	9.34
Rainfall	0.30	0.00	0.20	0.40

was the most important variable in the RF model followed by elevation, PI, Band 1, and Band 2. In contrast, the result of BRT indicated that Band 1 was the most important variable affecting SOC stocks, followed by PI, elevation, Band 2 and EVI. The majority of the topographic variables showed a very low contribution to the model, as did uranium concentration and terrestrial gamma-ray dose rate.

Due to the minor importance attributed to some of the predictor variables, we used the variables selected by stepwise regression (SR_RF and SR_BRT; total 11 predictors) to examine whether prediction accuracy would remain unchanged or increase when using a smaller number

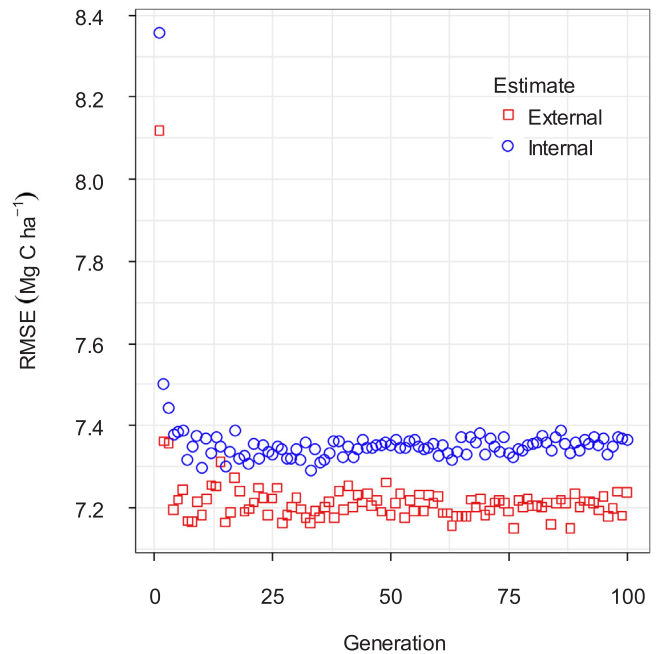


Fig. 6. Mean of external and internal RMSE calculated using 10-fold cross-validation with 5 repetitions for 100 iterations. The lowest external RMSE indicated the best iteration. In the final search using all data sets, the minimum external RMSE for SOC stocks at iteration 87 was $7.15 \text{ Mg C ha}^{-1}$ ($R^2 = 0.45$).

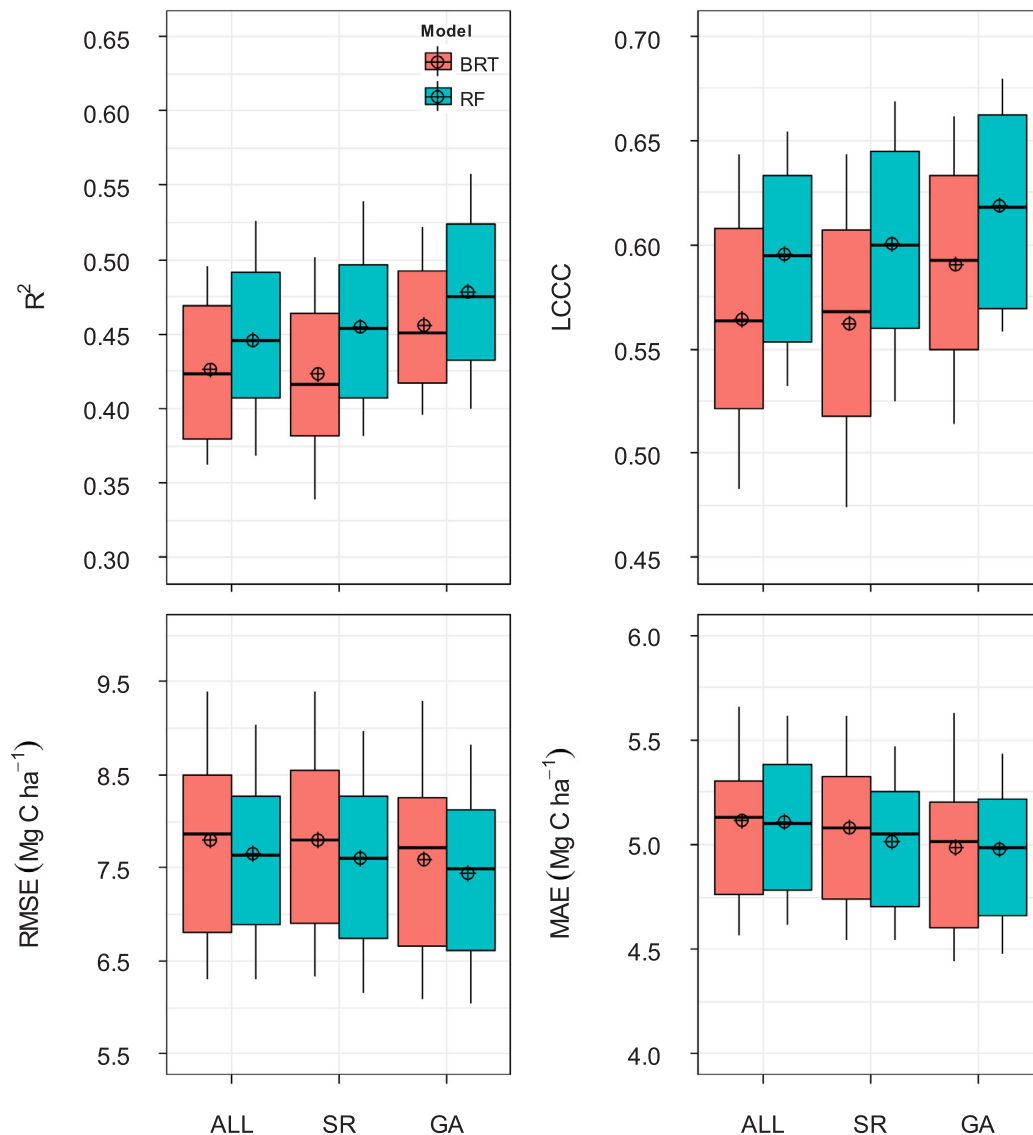


Fig. 7. Results of model evaluation criteria for prediction of soil organic carbon stocks (SOC, Mg C ha^{-1}) using boosted regression tree (BRT) and random forest (RF) models with 50 runs for the different input selections (ALL: 28 predictors; SR: 11 out of 28 predictors selected by Stepwise Regression method ($P < 0.05$); GA: 11 of 28 predictors selected by Genetic Algorithms). The coefficient of determination (R^2), Lin's concordance correlation coefficient (LCCC), root mean squared error (RMSE), and mean absolute error (MAE) are used to evaluate accuracy. The black lines within the box indicate the medians with 50 runs while crosshairs indicate means. Box boundaries indicate the 25th and 75th percentiles, whiskers below and above the box indicate the 10th and 90th percentiles.

of variables. The R^2 obtained by SR_RF increased compared to All_RF, and SR_RF performed better in terms of prediction error, indicating that a more parsimonious model did not impact its capability in predicting SOC stocks (Fig. 7). However, a smaller increase in accuracy was found in the BRT model. The order of the variables according to their relative contribution to the model did not change substantially, with elevation being the most important, followed by rainfall, Band 1, NBR and PI in the RF model (Fig. 8b). For the BRT model, the top five variables (in order of importance) were Band 1, PI, NBR, elevation and potassium concentration.

Applying the condensed dataset (in total 11 variables) selected by GA to RF resulted in the highest prediction accuracy with R^2 being equal to 0.48 ($\text{RMSE} = 7.44 \text{ Mg C ha}^{-1}$, $\text{MAE} = 4.98 \text{ Mg C ha}^{-1}$ and $\text{LCCC} = 0.62$). When GA_RF was compared with All_RF, the GA_RF model improved the predictive performance by increasing the R^2 and LCCC by 7.3% and 4.0%, respectively, while also reducing the RMSE

and MAE by 2.8% and 2.5% respectively. For the GA_RF model, rainfall, elevation, PI, Band 1 and MSAVI were the most important variables to predict SOC stock while TPI was identified as being of minor importance to SOC (Fig. 8c). Similarly, the BRT model also predominately identified Band 1, PI, MSAVI and elevation as the top four most important variables.

4. Discussion

In our study, the RF model outperformed the BRT model regardless of input features, expressed by larger R^2 and smaller prediction errors. This is in contrast to the results of Yang et al. (2016) who reported that BRT performed better than RF model in predicting SOC stocks on the Tibetan Plateau, China. This is likely to be due to differences in data sources, scale of prediction and types of predictor variables used. The superior performance of the RF model in terms of the prediction error

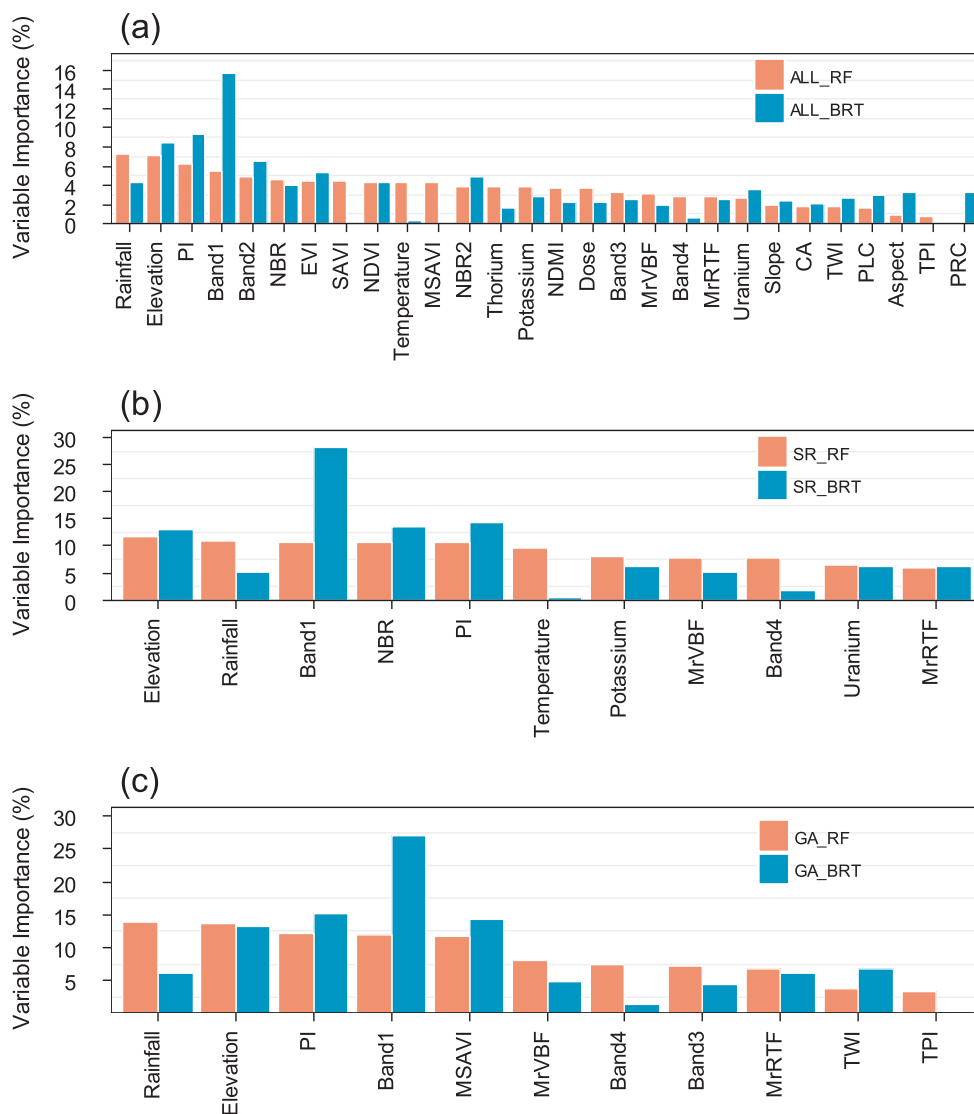


Fig. 8. Patterns in the importance of each predictor variable used in RF and BRT models to predict SOC stocks. Each variable was scaled to sum to 100%. (a) ALL: All 28 predictor variables; (b) SR: 11 out of 28 predictors selected by Stepwise Regression method and (c) GA: 11 of 28 predictors selected by Genetic Algorithms.

found in the present study is likely attributed to reduced susceptibility to over-fitting and better capacity to manage the hierarchical non-linear relationships that exist between SOC stocks and predictor variables. Several studies have reported that the performance of different tree-based models including RF and BRT can vary from study to study significantly, thus there is no general best modelling technique for predicting SOC stocks (Ließ et al., 2016; Were et al., 2015; Yang et al., 2016). Furthermore, it has been suggested that the predictive power of the modelled output is as much a result of the research design as the input variables (Aertsen et al., 2010). The results from this study demonstrated that RF based on GA feature selection is the most promising approach for faster, more efficient and better data interpretation and this result provides new insight into feature selection and model choice for estimating SOC stocks in semi-arid rangelands of eastern Australia.

The validation results (Fig. 7) show that the GA_RF model could explain variances of 48% for SOC stocks, which are comparable to recent studies predicting SOC stocks in semi-arid landscapes. For example, Wiesmeier et al. (2011) found the RF model combined with an assessment of variable importance (i.e. Classification and Regression Trees; RF_CRAT) could explain 53.4% of variation in the model building process in their study in Inner Mongolia, Northern China. In contrast, Román-Sánchez et al. (2016) also used RF but achieved much lower explained variance of 18% in Sierra Morena, Spain. Gray et al. (2015) used Cubist linear piecewise decision tree and multiple linear

regression models to examine the key soil-forming factors controlling SOC stocks. Their independent validation of the SOC stock predictions showed a LCCC value of up to 0.68, which were similar to the LCCC values of validation points ranging between 0.52 and 0.72 (Fig. 7) found in our study.

Feature selection is the process of removing irrelevant predictors from the pool of predictor variables to enhance a model's performance and capability (Miller et al., 2015). Stepwise regression analysis is useful to identify major environmental variables affecting SOC stocks (Kumar et al., 2012). It evaluates all environmental variables for their potential contributions to SOC variation and excludes variables not statistically significant ($P < 0.05$) in the multiple regression model (Oliveira et al., 2012). Procedures using stepwise regression are relatively efficient to compute, and have the advantage of using a small subset of least-correlated variables without losing a significant portion of the explanatory power of the data, thus minimizing the impacts of multi-collinearity (correlations among predictor variables) on the regression model (Huang and Townshend, 2003; Wang et al., 2015). Our results showed smaller improvements in these four statistical indices when features selected by SR were compared with all features. This is because the SR method makes a predictor assessment based on the predictor-response correlation and evaluates each predictor individually, thereby ignoring potentially important predictor interactions. However, the parsimonious model derived from GA yielded the

higher R^2 and LCCC with the smaller RMSE and MAE for both RF and BRT model compared with SR procedure. Our results showed that feature selection using GA enhanced the model's performance with the least number of informative predictor variables possible, which is consistent with the results of Xie et al. (2015) who reported that the GA variable selection technique can improve the accuracy of soil type recognition by 3–10%.

Our study demonstrated that remotely sensed vegetation indices were important predictors of SOC stocks (Fig. 8a). The type, density and age of vegetation influences the stock and distribution of SOC due to litter fall and root attributes (Akpa et al., 2016), the loss of C from soil via OM decomposition by regulating fluxes in soil temperature and moisture (Sharafatmandrad et al., 2010) and by protecting the soil surface from erosion (Facelli and Pickett, 1991). Favourable soil conditions such as high nutrient and water-holding capacity promote vegetation growth and therefore OM input to soil, as SOC, the major constituent of soil OM, is an important driver of cation exchange, nutrient cycling, infiltration and water holding capacity. These conditions are also influenced by topographic variables, with the lowest parts of the landscape having deeper soil with higher clay content and therefore inherently greater capacity to accumulate SOC. That is, soil with high nutrient and water holding capacity can support more plant growth compared with low fertility, sandy soils. This plant response to soil fertility leads to high spectral indices that reflect parts of the landscape with high SOC stocks (Zhao et al., 2014). We used 4 bands of fractional cover data and 7 vegetation indices, which had a strong relationship with each other (Fig. 3). Each of the fractional cover and vegetation indices had advantages in measuring vegetation vigour. However, our GA technique selected Band 1 (bare ground), Band 3 (non-green vegetation), Band 4 (fitting error) and MSAVI from these 11 indices as proxies, to estimate SOC stocks with climate and topographic variables. As expected, Band 3 and MSAVI were positively correlated to SOC stocks while the proportion of bare ground fraction (Band 1) had a negative correlation with SOC stocks (Fig. 3). The MSAVI is a soil adjusted vegetation index developed as a modification of the NDVI for areas with a high degree of exposed soil surface. However, the importance of Band 1 was higher than MSAVI in our modelling. Our study area was characterised by a large bare soil fraction (e.g. Band 1 ranged between 18.0% and 33.9% across study sites), therefore one possible explanation is that bare ground may be driving the loss of SOC via soil erosion more than vegetation growth is influencing the OM input due to the absence of vegetation.

NDVI and EVI have been used as a surrogate for vegetation parameters to explain the distribution of SOC stocks (Kunkel et al., 2011). However, they were not important predictor variables contributing to SOC stocks in this study. This is likely due to the limitations of using vegetation indices in arid environments, including the effects of exposed soil and the quantity of standing dead vegetation and litter on the spectral response (Maynard and Levi, 2017). It has been reported that green vegetation indices were not reliable when the proportion of bare soil is more than 20% (Sankey and Weber, 2009). In addition, for this present study we only used vegetation indices collected during the time period of field survey and seasonal fractional cover data from as close to the soil sampling time as achievable. It is possible that the short time frame from which data were collected may not sufficiently reflect the variability of vegetation indices. While Gasparri et al. (2010) indicated that the relationship between remote sensing data and vegetative biomass depended on the date of image acquisition, in many cases the stock of SOC reflects the historical pattern of vegetation growth and therefore vegetation indices over a short temporal window may give insufficient information about the OM inputs to the soil. For example, Wilson et al. (2017) found that multi-decadal (28 years) EVI (vegetation index) data significantly improved the prediction of SOC in a grassland environment with high spatial variability. Therefore, multi-temporal Landsat images including single band, derived vegetation indices and fractional cover data should be tested to determine whether they can

improve current methods in model performance of SOC stock prediction in the rangelands.

Numerous studies have highlighted the importance of climate in predicting SOC stocks (Table 1) due to its influence on net primary productivity and thus OM input. The climate in the current study area is characterised by mild winters and hot, dry summers, with rainfall maxima occurring in spring and winter. The combination of high temperatures and low rainfall in summer produces frequent seasonal droughts to which the local vegetation is adapted. However, annual mean temperature was not an important explanatory variable in this study probably due to minor variation (ranging between 18.0 and 20.5 °C) over the relatively small study area (Ottoy et al., 2017). On the contrary, annual mean rainfall was identified as the most important variable by the All_RF and GA_RF model (Fig. 8). Although rainfall had a small contribution to SOC stocks in BRT models, PI as a measure of water balance (significantly correlated with annual rainfall, Fig. 3) was the second most important variable in determining SOC stocks. This is consistent with previous studies which showed that soil moisture is essential for biomass productivity in semi-arid rangelands driving both above and below ground OM input to soil (Allen et al., 2013; Kunkel et al., 2011; Yang et al., 2008).

5. Conclusions

This study evaluated alternative methods to determine the most reliable and accurate model to predict SOC stocks in the surface 0.30 m of the soil using field collected SOC data and remotely-sensed variables in the semi-arid rangelands of eastern Australia. The results suggest random forest with genetic algorithm is the most promising model for predicting SOC stocks and that there is some opportunity to use remotely sensed vegetation indices to predict SOC stocks in this environment. The validation results also show that the prediction accuracy of the random forest with genetic algorithm model was acceptable with explained variances of 48% for SOC stocks, which were comparable to other studies where SOC stocks have been predicted in semi-arid rangelands. Our present study has demonstrated the importance of feature selection prior to predicting SOC stocks, although the BRT model was relatively insensitive to the input features selected by SR. The results have also shown that the bare ground fraction, elevation, PI and rainfall were important variables explaining the observed variability of SOC stocks in this semi-arid environment, and the contributions of the other environmental factors were less important. It should be noted that we included radiometrics to represent geology and soil features in our study, and more information on these variables is likely to strengthen the model and its interpretation. The approach proposed here can be extended in data-scarce areas (e.g. rangelands) to produce detailed information about SOC stocks. As such, the results of this study are of particular importance to provide a statistical and theoretical basis for producing digital SOC stocks maps and monitoring change in SOC based on readily available satellite products across rangelands.

Acknowledgements

Funding for this research was provided by NSW Treasury and NSW Department of Primary Industries. We are very grateful to two reviewers for their very thorough and helpful comments which greatly improved our manuscript.

References

- Adhikari, K., Hartemink, A.E., Minasny, B., Bou Kheir, R., Greve, M.B., Greve, M.H., 2014. Digital mapping of soil organic carbon contents and stocks in denmark. *PLOS ONE* 9, e105519.
- Aertsens, W., Kint, V., van Orshoven, J., Özkan, K., Muys, B., 2010. Comparison and ranking of different modelling techniques for prediction of site index in Mediterranean mountain forests. *Ecol. Model.* 221, 1119–1130.
- Akpa, S.I.C., Odeh, I.O.A., Bishop, T.F.A., Hartemink, A.E., Amapu, I.Y., 2016. Total soil

- organic carbon and carbon sequestration potential in Nigeria. *Geoderma* 271, 202–215.
- Allen, D.E., Pringle, M.J., Bray, S., Hall, T.J., O'Reagain, P.O., Phelps, D., Cobon, D.H., Bloesch, P.M., Dalal, R.C., 2013. What determines soil organic carbon stocks in the grazing lands of north-eastern Australia? *Soil Res.* 51, 695–706.
- Badgery, W.B., Simmons, A.T., Murphy, B.M., Rawson, A., Andersson, K.O., Lonergan, V.E., van de Ven, R., 2013. Relationship between environmental and land-use variables on soil carbon levels at the regional scale in central New South Wales, Australia. *Soil Res.* 51, 645–656.
- Badgery, W.B., Simmons, A.T., Murphy, B.W., Rawson, A., Andersson, K.O., Lonergan, V.E., 2014. The influence of land use and management on soil carbon levels for crop-pasture systems in Central New South Wales, Australia. *Agric. Ecosyst. Environ.* 196, 147–157.
- Behrens, T., Zhu, A.X., Schmidt, K., Scholten, T., 2010. Multi-scale digital terrain analysis and feature selection for digital soil mapping. *Geoderma* 155, 175–185.
- Bonfatti, B.R., Hartemink, A.E., Giasson, E., Tornquist, C.G., Adhikari, K., 2016. Digital mapping of soil carbon in a viticultural region of Southern Brazil. *Geoderma* 261, 204–221.
- Bou Kheir, R., Greve, M.H., Böcher, P.K., Greve, M.B., Larsen, R., McCloy, K., 2010. Predictive mapping of soil organic carbon in wet cultivated lands using classification-tree based models: the case study of Denmark. *J. Environ. Manage.* 91, 1150–1160.
- Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32.
- Chan, K.Y., McCoy, D., 2010. Soil carbon storage under perennial pastures in the mid-north coast of New South Wales, Australia. *Trop. Grassl. J.* 44, 184–191.
- Chan, K.Y., Oates, A., Li, G.D., Conyers, M.K., Prangnell, R.J., Poile, G., Liu, D.L., Barchia, I.M., 2010. Soil carbon stocks under different pastures and pasture management in the higher rainfall areas of south-eastern Australia. *Soil Res.* 48, 7–15.
- J.H. Dane C.G. Topp 2002. *Methods of Soil Analysis: Part 4 Physical Methods, Agronomy no. 9* Soil Science Society of America Madison, WI, United States of America.
- Davy, M.C., Koen, T.B., 2013. Variations in soil organic carbon for two soil types and six land uses in the Murray Catchment, New South Wales, Australia. *Soil Res.* 51, 631–644.
- Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. *J. Anim. Ecol.* 77, 802–813.
- Everingham, Y., Sexton, J., Skocaj, D., Inman-Bamber, G., 2016. Accurate prediction of sugarcane yield using a random forest algorithm. *Agron. Sust. Dev.* 36, 1–9.
- Facelli, J.M., Pickett, S.T.A., 1991. Plant litter: its dynamics and effects on plant community structure. *Botan. Rev.* 57, 1–32.
- Gallant, J., Austin, J., 2012. *Prescott Index Derived from 1" SRTM DEM-s. V2, CSIRO. Data Collection.* doi: 10.4225/08/53EB2D0EAE377.
- Gasparri, N.I., Parmuchi, M.G., Bono, J., Karszenbaum, H., Montenegro, C.L., 2010. Assessing multi-temporal Landsat 7 ETM+ images for estimating above-ground biomass in subtropical dry forests of Argentina. *J. Arid. Environ.* 74, 1262–1270.
- Gray, J.M., Bishop, T.F.A., Wilson, B.R., 2015. Factors controlling soil organic carbon stocks with depth in eastern Australia. *Soil Sci. Soc. Am. J.* 79, 1741–1751.
- Guo, P.-T., Li, M.-F., Luo, W., Tang, Q.-F., Liu, Z.-W., Lin, Z.-M., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. *Geoderma* 237–238, 49–59.
- Hengl, T., Heuvelink, G.B.M., Kempen, B., Leenaars, J.G.B., Walsh, M.G., Shepherd, K.D., Sila, A., MacMillan, R.A., Mendes de Jesus, J., Tamene, L., Tondoh, J.E., 2015. Mapping soil properties of Africa at 250 m resolution: random forests significantly improve current predictions. *PLOS ONE* 10, e0125814.
- Heung, B., Bulmer, C.E., Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. *Geoderma* 214–215, 141–154.
- Hobley, E., Wilson, B., Wilkie, A., Gray, J.M., Koen, T., 2015. Drivers of soil organic carbon storage and vertical distribution in Eastern Australia. *Plant Soil* 390, 111–127.
- Huang, C.-L., Wang, C.-J., 2006. A GA-based feature selection and parameters optimization for support vector machines. *Exp. Syst. Appl.* 31, 231–240.
- Huang, C., Townshend, J.R.G., 2003. A stepwise regression tree for nonlinear approximation: applications to estimating subpixel land cover. *Int. J. Rem. Sens.* 24, 75–90.
- IUSS-Working-Group, WRB, 2015. *World Reference Base for Soil Resources 2014, update 2015. International soil classification system for naming soils and creating legends for soil maps. World Soil Resources Reports No. 106.* FAO, Rome.
- Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environ. Modell. Softw.* 16, 309–330.
- Keating, B., Grundy, M., Battaglia, M., Eady, S., 2009. An analysis of greenhouse gas mitigation and carbon biosequestration opportunities from rural land use. *St Lucia, QLD: CSIRO; 2009. changeme:822.* doi: 10.4225/08/58615c9dd6942.
- Kuang, B., Tekin, Y., Mouazen, A.M., 2015. Comparison between artificial neural network and partial least squares for on-line visible and near infrared spectroscopy measurement of soil organic carbon, pH and clay content. *Soil and Tillage Research* 146, Part B, 243–252.
- Kuhn, M., 2008. *Building Predictive Models in R Using the caret Package.* 2008 28, 26.
- Kumar, S., Lal, R., Liu, D., 2012. A geographically weighted regression kriging approach for mapping soil organic carbon stock. *Geoderma* 189–190, 627–634.
- Kunkel, M.L., Flores, A.N., Smith, T.J., McNamara, J.P., Benner, S.G., 2011. A simplified approach for estimating soil carbon and nitrogen stocks in semi-arid complex terrain. *Geoderma* 165, 1–11.
- Ließ, M., Glaser, B., Huwe, B., 2012. Uncertainty in the spatial prediction of soil texture: comparison of regression tree and Random Forest models. *Geoderma* 170, 70–79.
- Ließ, M., Schmidt, J., Glaser, B., 2016. Improving the spatial prediction of soil organic carbon stocks in a complex tropical mountain landscape by methodological specifications in machine learning approaches. *PLOS ONE* 11, e0153673.
- Lima, A.R., Cannon, A.J., Hsieh, W.W., 2016. Forecasting daily streamflow using online sequential extreme learning machines. *J. Hydrol.* 537, 431–443.
- Maynard, J.J., Levi, M.R., 2017. Hyper-temporal remote sensing for digital soil mapping: characterizing soil-vegetation response to climatic variability. *Geoderma* 285, 94–109.
- Miller, B.A., Koszinski, S., Wehrhan, M., Sommer, M., 2015. Impact of multi-scale predictor selection for modeling soil properties. *Geoderma* 239–240, 97–106.
- Minasny, B., McBratney, A.B., Malone, B.P., Wheeler, I., 2013. Chapter one – digital mapping of soil carbon. In: Donald, L.S. (Ed.), *Advances in Agronomy.* Academic Press, pp. 1–47.
- Mouazen, A.M., Kuang, B., De Baerdemaeker, J., Ramon, H., 2010. Comparison among principal component, partial least squares and back propagation neural network analyses for accuracy of measurement of selected soil properties with visible and near infrared spectroscopy. *Geoderma* 158, 23–31.
- Muir, J., Schmidt, M., Tindall, D., Trevithick, R., Scarth, P., Stewart, J., 2011. Guidelines for Field measurement of fractional ground cover: a technical handbook supporting the Australian collaborative land use and management program. Tech. rep., Queensland Department of Environment and Resource Management for the Australian Bureau of Agricultural and Resource Economics and Sciences, Canberra.
- Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., Pereira, J.M.C., 2012. Modeling spatial patterns of fire occurrence in mediterranean Europe using multiple regression and random forest. *For. Ecol. Manage.* 275, 117–129.
- Orgill, S.E., Condon, J.R., Kirkby, C.A., Orchard, B.A., Conyers, M.K., Greene, R.S.B., Murphy, B.W., 2017a. Soil with high organic carbon concentration continues to sequester carbon with increasing carbon inputs. *Geoderma* 285, 151–163.
- Orgill, S.E., Waters, C.M., Melville, G., Toole, I., Alemseged, Y., Smith, W., 2017b. Sensitivity of soil organic carbon to grazing management in the semi-arid rangelands of south-eastern Australia. *Rangeland J.* 39, 153–167.
- Otto, S., De Vos, B., Sindayihebura, A., Hermy, M., Van Orshoven, J., 2017. Assessing soil organic carbon stocks under current and potential forest cover using digital soil mapping and spatial generalisation. *Ecol. Ind.* 77, 139–150.
- Page, K.L., Dalal, R.C., Pringle, M.J., Bell, M., Dang, Y.P., Radford, B., Bailey, K., 2013. Organic carbon stocks in cropping soils of Queensland, Australia, as affected by tillage management, climate, and soil characteristics. *Soil Res.* 51, 596–607.
- Poggio, L., Gimona, A., Brewer, M.J., 2013. Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. *Geoderma* 209–210, 1–14.
- Rabbi, S.M.F., Tighe, M., Cowie, A., Wilson, B.R., Schwenke, G., McLeod, M., Badgery, W., Baldock, J., 2014. The relationships between land uses, soil management practices, and soil carbon fractions in South Eastern Australia. *Agric. Ecosyst. Environ.* 197, 41–52.
- Rabbi, S.M.F., Tighe, M., Delgado-Baquerizo, M., Cowie, A., Robertson, F., Dalal, R., Page, K., Crawford, D., Wilson, B.R., Schwenke, G., McLeod, M., Badgery, W., Dang, Y.P., Bell, M., O'Leary, G., Liu, D.L., Baldock, J., 2015. Climate and soil properties limit the positive effects of land use reversion on carbon storage in Eastern Australia. *Sci. Rep.* 5, 17866.
- Rayment, G.E., Lyons, D., 2011. *Soil Chemical Methods-Australasia.* CSIRO Publishing, Collingwood, Victoria Australia.
- Razakamanarivo, R.H., Grinand, C., Razafindrakoto, M.A., Bernoux, M., Albrecht, A., 2011. Mapping organic carbon stocks in eucalyptus plantations of the central highlands of Madagascar: A multiple regression approach. *Geoderma* 162, 335–346.
- Robertson, F., Crawford, D., Partington, D., Oliver, I., Rees, D., Aumann, C., Armstrong, R., Perris, R., Davey, M., Moodie, M., Baldock, J., 2016. Soil organic carbon in cropping and pasture systems of Victoria, Australia. *Soil Res.* 54, 64–77.
- Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., Chica-Rivas, M., 2015. Machine learning predictive models for mineral prospectivity: an evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geol. Rev.* 71, 804–818.
- Rodriguez-Galiano, V.F., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P.M., Jeganathan, C., 2012. Random Forest classification of Mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Rem. Sens. Environ.* 121, 93–107.
- Román-Sánchez, A., Vanwalleghe, T., Peña, A., Laguna, A., Giráldez, J.V., 2016. Controls on soil carbon storage from topography and vegetation in a rocky, semi-arid landscapes. *Geoderma*.
- Sankey, T.T., Weber, K., 2009. *Rangeland Assessments Using Remote Sensing: Is NDVI Useful? Pages 113-122 in K.T. Weber and K. Davis (Eds.), Final Report: Comparing Effects of Management Practices on Rangeland Health with Geospatial Technologies (NNX06AE47G).* 168 pp.
- Schillaci, C., Acutis, M., Lombardo, L., Lipani, A., Fantappiè, M., Märker, M., Saia, S., 2017. Spatio-temporal topsoil organic carbon mapping of a semi-arid Mediterranean region: the role of land use, soil texture, topographic indices and the influence of remote sensing data to modelling. *Sci. Total Environ.* 601–602, 821–832.
- Schuman, G.E., Janzen, H.H., Herrick, J.E., 2002. Soil carbon dynamics and potential carbon sequestration by rangelands. *Environ. Pollut.* 116, 391–396.
- Schwenke, G.D., McLeod, M.K., Murphy, S.R., Harden, S., Cowie, A.L., Lonergan, V.E., 2013. The potential for sown tropical perennial grass pastures to improve soil organic carbon in the North-West Slopes and Plains of New South Wales. *Soil Res.* 51, 726–737.
- Sharafatmandrad, M., Mesdaghi, M., Bahremand, A., Barani, H., 2010. The role of litter in rainfall interception and maintenance of superficial soil water content in an arid rangeland in Khabr National Park in south-eastern Iran. *Arid Land Res. Manage.* 24, 213–222.
- Six, J., Conant, R.T., Paul, E.A., Paustian, K., 2002. Stabilization mechanisms of soil organic matter: Implications for C-saturation of soils. *Plant Soil* 241, 155–176.
- Sreenivas, K., Dadhwal, V.K., Kumar, S., Harsha, G.S., Mitran, T., Sujatha, G., Suresh, G.J.R., Fyze, M.A., Ravisankar, T., 2016. Digital mapping of soil organic and inorganic carbon status in India. *Geoderma* 269, 160–173.

- Stewart, C., Paustian, K., Conant, R., Plante, A., Six, J., 2007. Soil carbon saturation: concept, evidence and evaluation. *Biogeochemistry* 86, 19–31.
- Taghizadeh-Mehrjardi, R., Nabiollahi, K., Kerry, R., 2016. Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region. *Iran. Geoderma* 266, 98–110.
- Takata, Y., Funakawa, S., Akshalov, K., Ishida, N., Kosaki, T., 2007. Spatial prediction of soil organic matter in northern Kazakhstan based on topographic and vegetation information. *Soil Sci. Plant Nutrit.* 53, 289–299.
- Thompson, J.A., Pena-Yewtukhiw, E.M., Grove, J.H., 2006. Soil–landscape modeling across a physiographic region: Topographic patterns and model transportability. *Geoderma* 133, 57–70.
- Vågen, T.-G., Winowiecki, L.A., Tondoh, J.E., Desta, L.T., Gumbricht, T., 2016. Mapping of soil properties and land degradation risk in Africa using MODIS reflectance. *Geoderma* 263, 216–225.
- Wang, B., Chen, C., Liu, D.L., Asseng, S., Yu, Q., Yang, X., 2015. Effects of climate trends and variability on wheat yield variability in eastern Australia. *Clim. Res.* 64, 173–186.
- Waters, C.M., Melville, G.J., Orgill, S.E., Alemseged, Y., 2015. The relationship between soil organic carbon and soil surface characteristics in the semi-arid rangelands of southern Australia. *Rangeland J.* 37, 297–307.
- Waters, C.M., Orgill, S.E., Melville, G.J., Toole, I.D., Smith, W.J., 2016. Management of grazing intensity in the semi-arid rangelands of southern Australia: effects on soil and biodiversity. *Land Degrad. Dev.* 28, 1363–1375.
- Welikala, R.A., Fraz, M.M., Dehmeshki, J., Hoppe, A., Tah, V., Mann, S., Williamson, T.H., Barman, S.A., 2015. Genetic algorithm based feature selection combined with dual classification for the automated detection of proliferative diabetic retinopathy. *Comput. Med. Imaging Graph.* 43, 64–77.
- Were, K., Bui, D.T., Dick, Ø.B., Singh, B.R., 2015. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afriomontane landscape. *Ecol. Ind.* 52, 394–403.
- Wiesmeier, M., Barthold, F., Blank, B., Kögel-Knabner, I., 2011. Digital mapping of soil organic matter stocks using random forest modeling in a semi-arid steppe ecosystem. *Plant Soil* 340, 7–24.
- Wilson, B.R., Koen, T.B., Barnes, P., Ghosh, S., King, D., 2011. Soil carbon and related soil properties along a soil type and land-use intensity gradient, New South Wales Australia. *Soil Use Manage.* 27, 437–447.
- Wilson, C.H., Caughlin, T.T., Rifai, S.W., Boughton, E.H., Mack, M.C., Flory, S.L., 2017. Multi-decadal time series of remotely sensed vegetation improves prediction of soil carbon in a subtropical grassland. *Ecol. Appl.* 27, 1646–1656.
- Xie, H., Zhao, J., Wang, Q., Sui, Y., Wang, J., Yang, X., Zhang, X., Liang, C., 2015. Soil type recognition as improved by genetic algorithm-based variable selection using near infrared spectroscopy and partial least squares discriminant analysis. *Sci. Rep.* 5, 10930.
- Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G., Li, D.-C., 2016. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecol. Ind.* 60, 870–878.
- Yang, Y., Fang, J., Tang, Y., Ji, C., Zheng, C., He, J., Zhu, B., 2008. Storage, patterns and controls of soil organic carbon in the Tibetan grasslands. *Glob. Change Biol.* 14, 1592–1599.
- Young, R., Wilson, B.R., McLeod, M., Alston, C., 2005. Carbon storage in the soils and vegetation of contrasting land uses in northern New South Wales. *Australia. Aust. J. Soil Res.* 43, 21–31.
- Zhao, M.-S., Rossiter, D.G., Li, D.-C., Zhao, Y.-G., Liu, F., Zhang, G.-L., 2014. Mapping soil organic matter in low-relief areas based on land surface diurnal temperature difference and a vegetation index. *Ecol. Ind.* 39, 120–133.